Whole-Body Detection, Identification and Recognition at Altitude and Range

Siyuan Huang, Ram Prabhakar Kathirvel, Yuxiang Guo, *Graduate Student Member, IEEE*, Chun Pong Lau, and Rama Chellappa, *Life Fellow, IEEE*

***Abstract*—In this paper, we address the challenging task of whole-body biometric detection, recognition, and identification at distances of up to 500m and large pitch angles of up to 50°. We present an end-to-end system evaluated on the challenging Biometric Recognition and Identification at Altitude and Range (BRIAR) dataset. Our approach involves pre-training the detector on common image datasets and fine-tuning it on BRIAR's complex videos and images. After detection, we extract body images and employ a feature extractor for recognition. We conduct thorough evaluations under various conditions, such as different ranges and angles in indoor, outdoor, and aerial scenarios. Our method achieves an average F1 score of 98.29% at IoU = 0.7 and demonstrates strong performance in recognition accuracy and true acceptance rate at low false acceptance rates compared to existing models. On a test set of 100 subjects with 444 distractors, our model achieves a rank-20 recognition accuracy of 75.13% and a TAR@1%FAR of 54.09%.**

***Index Terms*—Body recognition, long-range biometric identification, deep learning for biometric identification.**

I. Introduction

SINCE the introduction of YOLO [1], R-CNN [2], and ResNet [3], object/person detection, and identification based on deep learning has received extensive attention [4], [5], [6], [7], [8], [9]. While body detection shares many features of the general object detection task, it has additional challenges due to appearance changes, variations in scale, acquisition conditions, and background clutter [10]. Body detection is useful in various applications, including autonomous driving [11], search and rescue [12], [13], and verification and recognition [14] based on biometric information.

Compared with general object detection datasets (such as COCO [15]), a key challenge of body detection task is processing data acquired by cameras located at different angles, altitudes, locations, and ranges. Fig. 1 shows body images acquired by cameras located at multiple ranges and corresponding detection results. In some situations, human bodies are even incomplete.

Although there is a vast number of works based on various traditional algorithms [17], [18], [19], [20], [21], [22], [23] to detect bodies, they do not perform well on data collected at range and altitude. Deep learning methods [24], [25], [26], [27], [28], [29], [30] can extract better features and improve detection precision. While deep learning methods have almost monopolized the area of object detection, they are less explored for body detection. For example, among existing methods, only a few [31], [32] have explored body detection at high altitudes and long ranges.

Motivated by the aforementioned challenges, we present an end-to-end approach that can detect, identify and recognize human bodies at ranges up to 500m and pitch angles up to 50°. Our method requires the model to be pre-trained on a small number of public datasets and then fine-tuned on the BRIAR dataset. In both processes, the first stage of the model extracts features and generates proposals. In the second stage, the model detects the body based on these proposals. Finally, the model generates features from body images cropped from the detector. Our main contributions are:

- Our method synthesizes the features of public datasets and BRIAR data so that the model can obtain good representations without extensive training on large public datasets or BRIAR.
- We explore body detection, identification and recognition in detail at different altitudes, angles, scenes, and ranges. The proposed method can detect, identify and recognize human bodies under all these challenges with high accuracy.
- Experiments show that our model can maintain an F1 score of > 0.98 for BRIAR in almost all cases under different IoU thresholds, as well as a rank-20 accuracy of 90.36% and a TAR@1%FAR of 49.26% on the test set with a large number of distractors, indicating that the model is effective, robust, and stable at different ranges and angles.

II. Related Work

A. Body Detection

HOG [19] is one of the popular methods for body detection, based on RGB and optical flow. Many follow-on works [18],

Received 18 March 2024; revised 28 July 2024 and 13 September 2024; accepted 21 October 2024. Date of publication 28 October 2024; date of current version 27 June 2025. This work was supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) under Grant 2022-21102100005. This article was recommended for publication by Associate Editor J. Lu upon evaluation of the reviewers' comments. *(Corresponding author: Rama Chellappa.)*

Siyuan Huang, Ram Prabhakar Kathirvel, and Yuxiang Guo are with the Whiting School of Engineering, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: shuan124@jhu.edu; rprabha3@jhu.edu).

Chun Pong Lau is with the School of Data Science, City University of Hong Kong, Hong Kong (e-mail: cplau27@cityu.edu.hk).

Rama Chellappa is with the Whiting School of Engineering and the School of Medicine, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: rchella4@jhu.edu).

Digital Object Identifier 10.1109/TBIOM.2024.3487545

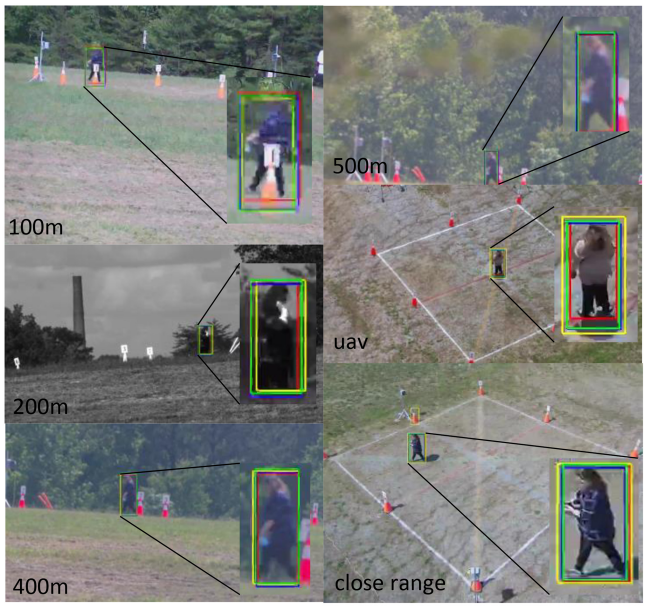


Fig. 1. Qualitative body detection results on the challenging BRIAR dataset [16]. Compared with general datasets, the BRIAR data has many challenges such as blur, occlusion, camera motion, and incomplete human body. Left: original BRIAR datasets at altitudes and ranges. Right: Detection results from our models. Red: ground truth. Other colors: detection results from different models. All subjects consented to be published.

[20], [21], [23], [33] also used HOG to detect bodies. Kamal and Jalal [34] detected occluded bodies using background subtraction. Khan et al. [35] used templates and 3D modeling multi-person detection and occlusion. Liu et al. [36] solved the problem of lighting and low-quality images by cascading head and shoulder detection of HOG features. Most of these methods did not take into account body movement, whereas each subject in BRIAR would continuously move in difficult scenes or stand still over a small number of frames.

Kim and Moon [24] utilized a CNN to directly extract features and classify videos. Ouyang et al. [26] used a multi-layer Restricted Boltzmann Machine (RBM) to extract features and detect bodies at multiple levels, addressing challenges due to occlusion and shadows. Zhang et al. [27] used a Faster R-CNN to achieve more accurate and specialized body system. Li et al. [28] replaced the CNN of Faster R-CNN by a dilated CNN to achieve body detection at a lower resolution. Li et al. [37] and Du et al. [29] fused features extracted by multiple R-CNN and SSD, respectively, to detect bodies at arbitrary scales. However, these methods have not been tested on datasets collected at high altitudes and long ranges.

B. Recognition and Identification

Image and video-based methods for person identification and recognition have been developed over the years. Several image-based models [38], [39], [40], [41] have shown impressive results on image-based benchmarks. More complex is the video-based task, which still requires significant research [42], [43], [44], [45], [46], [47], [48], [49], [50]. However, these methods did not provide any solutions for the complex acquisition conditions mentioned above, and their performance significantly deteriorates at long ranges and high altitudes. While some works addressed various difficulties, such as different angles [51], [52], different poses [53], [54], [55], and occlusions [56], [57], [58], [59], [60], [61], these methods were not effective when clothing variations are present.

Recently, researchers focusing on human recognition have paid attention to variations in clothing. These efforts mainly fall into two categories: extracting clothing-independent features from RGB images [62], [63], [64], [65] and using various multimodal information [66], [67], [68], [69], [70], [71] to extract features robust to variations in clothing. In the former, the latest model is Gu et al.'s CAL [43], which used a clothes-based adversarial loss to prevent the clothing classifier from identifying the same person wearing different attire, thereby extracting features that are not related to clothing.

In [72], the authors proposed a novel method comprising of three key modules: a clothing-attention module to mitigate clothing variations, a human semantic attention module to enhance semantic details, and an identity enhancement module focused on emphasizing pedestrian identity importance. Han et al. [73] introduced an augmentation strategy using a clothing-change covariance estimation method to identify significant changes in clothing. They employed adversarial learning to train an augmentation generator, ensuring effective augmentation while minimizing identity alteration. Additionally, they proposed an ID-correlated augmentation strategy to increase intra-ID clothing variations and decrease inter-ID variations. Yang et al. [74] introduced a causality-based auto-intervention model that separates clothing bias from identity representation using a dual-branch network. Yang et al. [75] proposed SirNet, a network to learn distinct feature embeddings from random samples without employing hard mining strategies. In contrast to existing methods, SirNet represents a sample from each identity as a cluster of points. Additionally, the authors suggested a feature augmentation technique to synthesize challenging samples.

Nguyen et al. [76] presented a method to extract body shape cues using a Relational Shape Embedding branch and train the network with contrastive viewpoint-aware loss (CVL). The CVL aligns body shape feature embeddings under varying viewpoints and enhances the discriminative capabilities of appearance embeddings across different identities and viewpoints. In [77], Huang et al. leveraged meta-learning to address the domain shift between training and test sets due to changes in clothing style. They proposed dividing the training set into meta-train and meta-test subsets and simulating different proportions of cloth-changing and cloth-consistent image pairs. A calibration loss was used to reduce disparity between cloth-changing and cloth-consistent types, and a ranking loss enhanced the approach's robustness.

In the realm of multi-modality, Jin et al. [68] used gait information to identify the body, while Arkushin et al. [78] used the face to identify the body. They have all achieved state-of-the-art performance on LTCC at the time of publication. However, they still have not solved the problem of directly identifying persons from raw, complex video data or addressing challenges encountered in real videos, such as height, long distances, and low resolution. Therefore, in our work,

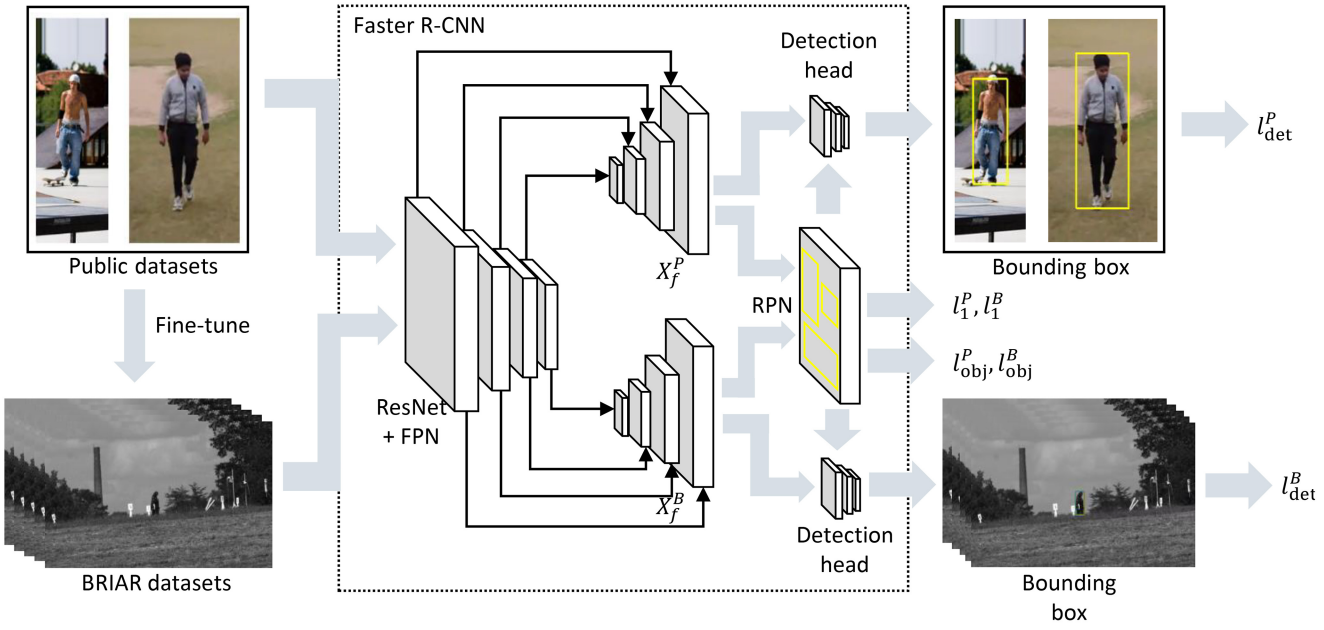


Fig. 2. The body detector pipeline consists of two stages: pre-training on public datasets and finetuning on the BRIAR dataset to learn a general semantic representation. Once the detector is trained, it is used to generate cropped body images from raw videos, which are utilized for recognition and identification.

we aim to simultaneously consider these challenges to achieve good recognition performance across the varied acquisition scenarios.

III. Methods

The goal of this work is to implement robust detection and recognition systems for whole-body-based biometrics using datasets collected at different altitudes and ranges. As shown in Fig. 3, our proposed BRIARNet model consists of a detector and a feature extractor stage. The training process is outlined as follows: *a)* Pre-training the detector on public datasets, *b)* Fine-tuning the detector on the BRIAR dataset, and finally *c)* Training the feature extractor on the BRIAR dataset. During inference, given a frame, the detector network (Fig. 2) identifies the location of the person and outputs a bounding box. The cropped body images are then passed through the feature extractor (Fig. 3) to produce semantic features, which are subsequently used for identification.

Detector pre-training: Fig. 2 shows the architecture of the person detector model. Specifically, it is based on Faster R-CNN [6] and ResNet-50, and learns the common features of public datasets and the BRIAR dataset. Initially, we pre-train the detector on publicly available datasets, COCO [15] and Visym [79], to obtain initial body feature information (more details are provided in Section IV-A). Given an input $X^P \in \mathbb{R}^{C\times H\times W}$ which is sampled from public datasets, where C is the channel size, H and W are height and width respectively, we extract features X_f^P from it with a feature extractor f. X_f^P can be multiple layers when the Feature Pyramid Networks (FPN) [8] is used. Then, we use Region Proposal Networks (RPN) [6] to calculate a proposal $b^P \in \mathbb{R}^4$ and its objectness score $o^P \in \mathbb{R}^2$ for X_f^P. b^P and o^P each have a loss. For b^P, RPN calculates its smoothed ℓ_1 loss [5] as

$$l_1 = \begin{cases} \frac{\left(b^P - \hat{b}^P\right)}{2\beta}, & \text{if } |b^P - \hat{b}^P| < \beta \\ |b^P - \hat{b}^P| - \frac{\beta}{2}, & \text{otherwise} \end{cases} \tag{1}$$

where $\hat{b}^P$ is the ground truth bounding box, β is smooth threshold. For o^P, RPN calculates its binary cross entropy

$$l_{\text{obj}} = -\left[\hat{o}^P \log o^P + \left(1 - \hat{o}^P\right) \log\left(1 - o^P\right)\right] \tag{2}$$

where $\hat{o}^P$ is the label of whether there is an object in the current proposal. Finally, the ROI head calculates the probability that the proposal belongs to a certain class. Since we target the specific class of body, we only have two classes in our approach, person and background. This part is also updated with binary cross entropy loss l_{det}. The total loss is

$$\mathcal{L} = l_1 + l_{\text{obj}} + l_{\text{det}} \tag{3}$$

Detector finetuning: After pre-training on public datasets, we fine-tune the model on the BRIAR dataset. For BRIAR's input X^B, the fine-tuning process is the same as was used for the public dataset. We train on multiple public and BRIAR datasets to obtain more synthesized features. However, the sizes of different datasets are quite different. In order to allow the model to sample more evenly, especially focusing on datasets with small sizes, a sampling strategy is employed. Assuming that dataset d_i has $|d_i|$ images or videos, the probability that a certain image or video is sampled from this dataset is $\frac{1}{|d_i|}$. Its final overall sampling probability is normalized on all datasets. Since the size of BRIAR is small at some ranges (e.g., 500m), this is crucial to improving the overall performance on the BRIAR data. Each video is sampled and then a fixed number of frames is uniformly taken

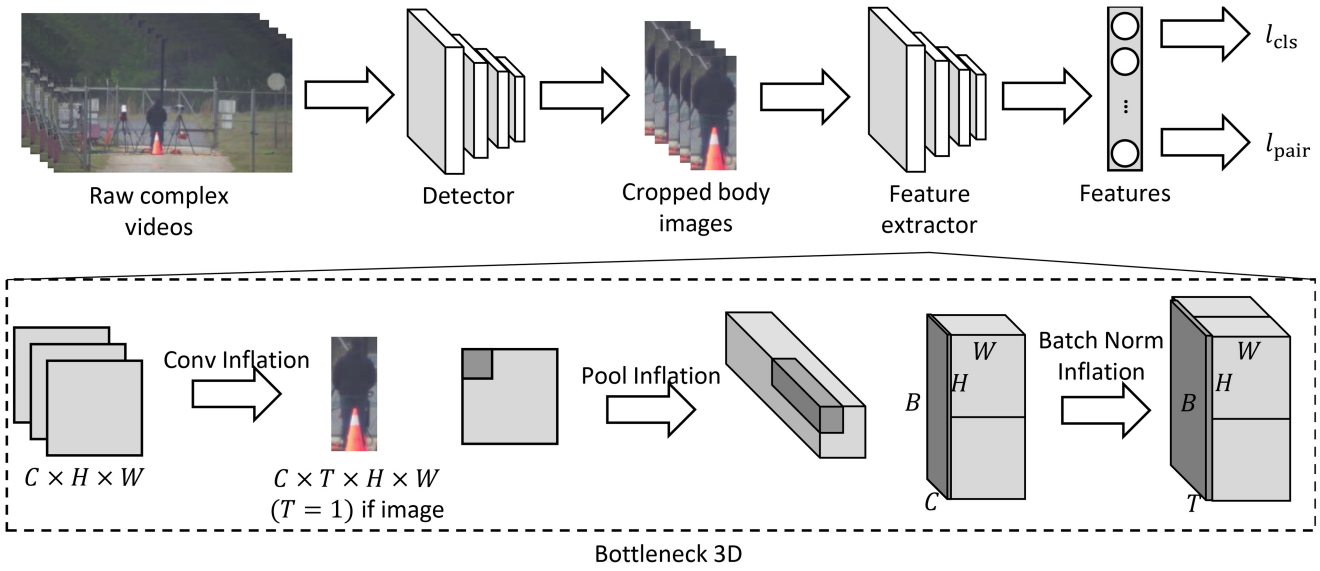


Fig. 3. The pipeline of the recognition model. We leverage an inclusion strategy to make sure the model can deal with images and videos simultaneously.

to form X. All X go through a series of augmentation before fed to the feature extractor.

Feature extractor training: Fig. 3 shows the pipeline of the recognition model. We use ResNet-50 [3] with 3D convolution blocks as the backbone network. Unlike the detector, we train the feature extractor network from scratch using only the BRIAR dataset (See Section IV-B for more details on the network architecture and training sampling strategy). The input to the recognition model is $X_{\text{Rec}} \in \mathbb{R}^{B\times C\times T\times H\times W}$, which is the body image sequence cropped from the detector, where T is the length of the time dimension. We adopt a strategy of sampling n people in each batch and k videos for each person, so that $B = nk$. On the other hand, since videos are often longer ($\sim 2k$ to $3k$ frames per video), it is computationally inefficient to directly load all the frames of each video. Furthermore, there is a lot of redundant information in successive frames. Therefore, we take one frame for every fixed stride. Since the length of each video is inconsistent, in order to ensure that all video frames in a batch are of the same length, we also record the index of valid frames of each video. If the length of a video exceeds T, we directly take a random frame, and feed consecutive T frames to the extractor. Otherwise, we zero pad to make it T. Since each person has controlled images of simple situations (such as images of standing indoors), in addition to complex videos, we also include these images during training. These images are equivalent to a video with a length of 1, so we zero pad them. The extractor ultimately outputs $X_f^{\text{REC}} \in \mathbb{R}^{B\times d}$ features, where d is the feature dimension. Finally, we calculate the l_{cls} and l_{pair} of the feature to achieve re-identification, where l_{cls} is the cross entropy of subjects, and l_{pair} is the triplet loss [80]. The final loss is,

$$\mathcal{L}_{\text{rec}} = l_{\text{cls}} + l_{\text{pair}} \tag{4}$$

We propose an inclusion strategy to enable a unified model capable of processing both image and video inputs. This strategy involves extracting temporal features when handling video inputs, while ensuring independence of the time dimension for 2D images. Initially, a batch of images is conceptualized as a batch of videos with each video containing a single frame. This approach allows us to employ a consistent architecture for handling both spatial and temporal features. When processing image batches, where the temporal dimension T is 1, temporal operations are bypassed, focusing solely on spatial operations. In contrast, video batches undergo both temporal and spatial operations. Consequently, our model guarantees that image inputs retain only spatial information, while video inputs incorporate both spatial and temporal information.

It is important to note that our inclusion strategy is not tied to specific layers or modules, but rather a flexible approach applicable to various layers and modules within neural network architectures. In our experiments, we applied the inclusion strategy to convolutional, pooling, and batch normalization layers, demonstrating its versatility and effectiveness across different components of the model.

IV. Experiments

A. Datasets

BRIAR: The BRIAR dataset [16] consists of images and videos of people moving in various situations, and is divided into face and whole body. The entire dataset has $> 350,000$ images and $1,300$ hours of videos. A total of 1000 subjects participated in the production of the entire dataset. Specifically, the BRIAR data can be divided into multiple datasets according to different altitudes, ranges, environments, and actions. The BRIAR data has three environments: indoor, outdoor and aerial (Fig. 4). The camera in the indoor environment is in a relatively fixed position, and various interferences will be relatively small. Therefore, indoor images and videos are relatively good for training. The indoor subjects have two kinds of movements. First, each subject walks along a preset and then walks randomly in the same area. These movement patterns constitute two datasets, *struct* and *rand*. In addition

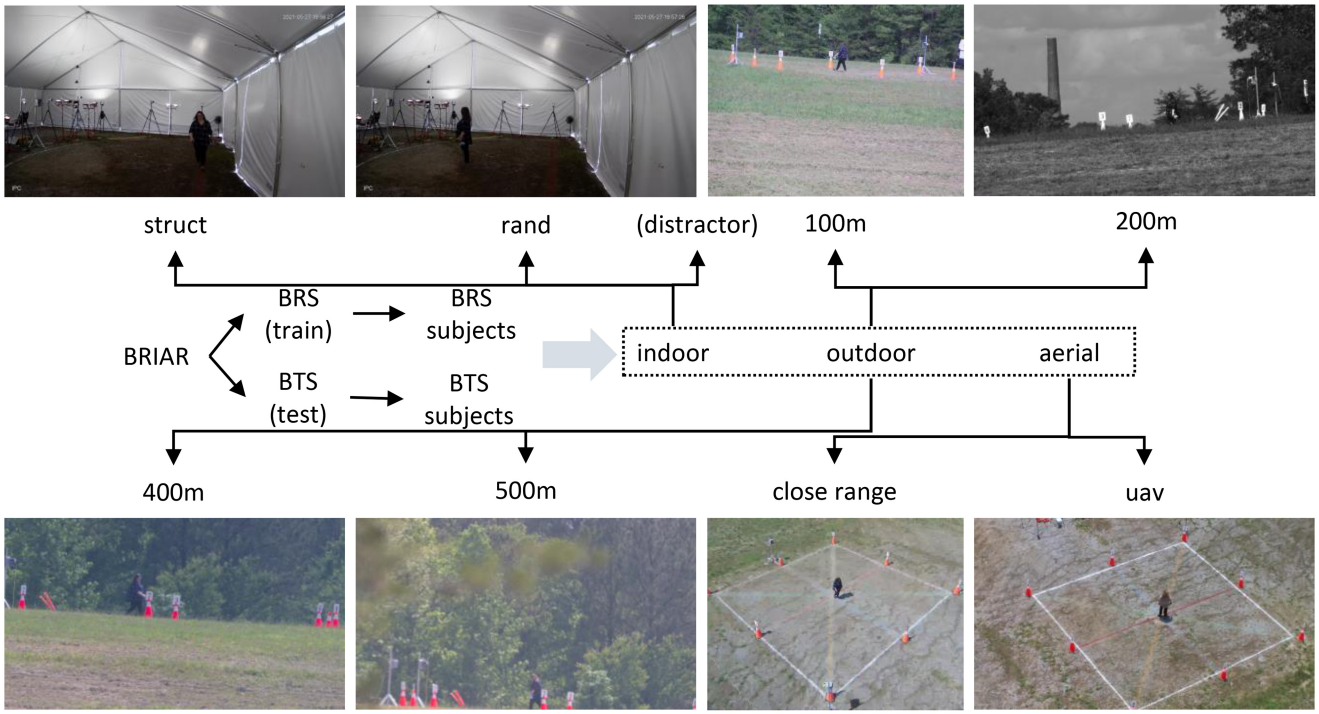


Fig. 4. BRIAR's structure tree. Each experimental field has some colored lines. *struct* requires subjects to walk along these lines, while in *rand* subjects walk freely. Note that the data for outdoor environments is the most complex. All subjects are smaller due to acquisition at large range. 200m videos pose challenges due to occlusion. A considerable portion of 500m subjects is at the very bottom or even the corner of frames. Furthermore, a large number of outdoor videos suffer from camera shake regardless of ranges, and the shake increases significantly as camera range increases. These issues make robust detection on BRIAR data challenging. All images shown have subjects who have given consent.

to the normally collected data, there is also a dedicated indoor distractor dataset. This dataset is full of images that are processed under various perturbations or scenarios that differ from the training set, which is equivalent to "physical" augmentation.

Datasets for the outdoor environment are more complex. According to different camera ranges, the outdoor data is divided into four datasets: *100m*, *200m*, *400m*, and *500m*. Since interferences such as outdoor ambient light and occlusion are much stronger in outdoors, these datasets usually challenge the detector. The quality of outdoor videos is significantly lower than indoor videos (Fig. 4). Therefore, our sampling strategy can focus more on feature learning in difficult outdoor environments. Finally, for aerial data, based on different altitudes and angles, BRIAR data can be divided into *close range* and Unmanned Aerial Vehicles (*UAV*), where close range represents aerial videos near the ground, and UAV represents aerial videos with higher altitudes. Both outdoor and aerial data have struct and rand actions. Fig. 4 shows the structure of BRIAR data as a tree.

The annotation information for BRIAR data includes various attributes of subjects such as birth date, gender, weight, etc. and ground truth bounding boxes. In our experiments, we only need the bounding boxes. We used data from the whole body part. The details on all datasets of BRIAR we use in this paper is shown in Table I.

COCO and Visym: For public datasets, we mainly use COCO [15] and Visym [79]. COCO is one of the most classic datasets for object detection tasks, with the advantages of large size and various classes. Visym is a large public dataset

TABLE I
THE TRAINING AND TEST SET VIDEOS OF BRIAR DATASETS USED IN THIS PAPER

Dataset	Training	Test
struct	2,268	1,225
rand	2,257	1,208
100m	729	382
200m	887	486
400m	767	351
500m	695	407
close range	3,553	1,843
UAV	119	75

dedicated to people, with millions of video clips of hundreds of human activities. Since one of our goals is to reduce the training costs while at the same time learn the representations of BRIAR datasets faster, we only extract very few images from COCO and Visym (at least two orders of magnitudes smaller than the full datasets). Specifically, for COCO, we selected 567 and 64 images with clear bodies as training and test sets, respectively. For some of the images, we manually re-labeled the bounding box. For Visym, we selected 10,430 and 1,064 images from videos of people walking as training and test sets. Section IV-C shows that with just this of publicly available datasets, our fine-tuned model's results on BRIAR are comparable to pre-trained models trained on full COCO.

B. Model Settings

We mainly use Faster R-CNN [6] based on three model settings: pre-trained (PRE), fine-tuned (FT), and from scratch

(SCR). Among them, the pre-trained model refers to the model trained only on public datasets we choose, without any BRIAR information. The fine-tuned model refers to fine-tuning on the BRIAR dataset after the previous pre-trained model is trained, and the scratch model refers to a model that is directly trained on the BRIAR dataset without using the public dataset. We compare all models to pre-trained Detectron 2 (DET2) [81]. During training, all models first sample videos according to the sampling strategy discussed in Section III, and then randomly select five frames from each video. At test time, for each video in the test set, we sample frames at fixed time intervals. Therefore, the number of frames taken by each video is different. In order to balance the number of frames taken for each dataset and speed up the test, time intervals in different datasets are set at different values. For struct, rand, and close range, we take one frame for every 300 frames (∼10s, if FPS = 30). For other datasets, models take one frame every 150 frames. Since the BRIAR data is about 1 - 2 minutes per video, at the end about 10 - 20 frames remain. We use the F1 score as the final evaluation metric. Data augmentations include grayscale ($p = 0.25$), brightness, saturation, contrast, horizontal flip ($p = 0.5$), and RGB shift ($p = 0.3$, limit $= \pm 30$ RGB values).

The body detector of our model is based on ResNet-50 [3] and the 5-layer FPN [8]. The β of ℓ_1 loss is $\frac{1}{9}$. The optimizer is SGD with momentum 0.9. The initial learning rate is 0.00001. There is ℓ_2 regularization with a decay rate $= 0.0001$. The pre-trained model needed about 11 min/epoch on 8x NVIDIA RTX A5000 GPUs. Fine-tuned and scratched models took about 3 hours per epoch on 7× GeForce RTX 3090 or 6× A5000 GPUs. The fine-tuned model was finally trained for ten epochs, while the scratch model was trained for twenty-five epochs. A batch size of 1 was used for each GPU. We also compare performances with the Faster R-CNN implementation of Detectron2 (DET2) [81] based on complete COCO training. We adopted the model that was closest to the structure of our current model.

The recognition model we used is built using ResNet-50 with 3D convolution blocks. We used the optimizer Adam with an initial learning rate of 1×10^{-4}. Additionally, we applied ℓ_2 regularizer with a decay rate of 5×10^{-4}. The batch size is 16, which corresponds to 4 people ×4 videos per person.

C. Results

1) Detection in Indoor: We discuss the performance of each model under different altitudes and ranges according to the situation. Table II shows the F1 score of all models in indoor struct and rand under different IoU thresholds. Since the indoor situation is the simplest in BRIAR and closest to the situation of public datasets, this comparison can directly reflect the preliminary performance of each model. First, under all IoU thresholds, fine-tuned, scratch, and Detectron2 models achieve almost similar and excellent performance (> 98% of F1 score), while the pre-trained model is not as good as the other three, although it also reached > 90% of F1. Since the pre-trained model has no BRIAR information and only has a small part of public datasets, even some of the simplest indoor cases are not detected. Correspondingly, the Detectron2 trained on the complete COCO achieves similar performance to fine-tuned and scratched models. This basically shows that for datasets with more variations, either specific training for its features, or training on large public datasets to obtain an overall feature distribution, is required.

TABLE II
F1 SCORE COMPARISON IN INDOOR DATASETS OF BRIAR UNDER DIFFERENT IOU THRESHOLDS. LOW IOU WILL HELP MODELS DETECT BODIES, BECAUSE NOW A PROPOSAL IS MORE LIKELY TO BE ASSIGNED AS A TRUE BOUNDING BOX. HOWEVER, IT WILL ALSO INCREASE THE FALSE POSITIVE RATE BECAUSE NOW ANY PROPOSAL IS MORE LIKELY TO BE ASSIGNED A TRUE BOUNDING BOX. FOR HIGH IOU, THE PHENOMENON IS THE OPPOSITE

	0.35		0.5		0.7	
model	**struct**	**rand**	**struct**	**rand**	**struct**	**rand**
PRE	91.84	97.15	91.76	97.07	86.34	94.43
FT	99.13	99.82	99.11	99.81	98.73	99.60
SCR	99.17	99.91	99.17	99.91	98.80	99.85
DET2	98.77	99.94	98.77	99.94	98.67	99.91

On the other hand, the performances of the fine-tuned, scratched, and Detectron2 models are very stable and do not change much with IoU. This shows that the bounding boxes predicted by these three models are relatively close to the ground truth. In contrast, the performance of the pre-trained model drops significantly under high IoU thresholds, which shows that its prediction is only a rough estimate, and the body has not been accurately localized. Furthermore, all models perform better on rand than struct. We believe this may be because in the rand situation people show multiple angles in front of cameras, while in struct mode view angles are very limited. Such differences may cause models to better recognize rand patterns (see Fig. 5 for comparison).

2) Detection in Outdoor: Table III shows the F1 score of all models in different outdoor ranges under different IoU thresholds. First, the three models of fine-tuned, scratch, and Detectron2 have similar, excellent, and stable performance (> 98% of F1) in the cases of 100m, 200m, and 400m, while the performance of the pre-trained model in outdoor is significantly lower than for indoor. This validates the conclusions found in indoor datasets. Second, except for 100m, all models in other ranges show that the performance decreases as the range goes farther. The performance of models at 200m is better than that of 100m. This may be because grayscale helps models to distinguish distinct frames, although it also blends the body and background in few frames. In addition, as the IoU threshold increases, the performance of models at 100m is significantly more stable than that at 400m and 500m, which is in line with the findings between interference and range mentioned above.

When the IoU threshold is not high, the fine-tuned model and scratch model are better than Detectron2 at 200m and 400m. But when the IoU threshold is high, the former is worse than the latter. This shows that models with BRIAR information tend to find the matching pattern first, and then consider localization, while models without BRIAR information behave just the opposite. This difference strongly illustrates the influence of BRIAR information on models

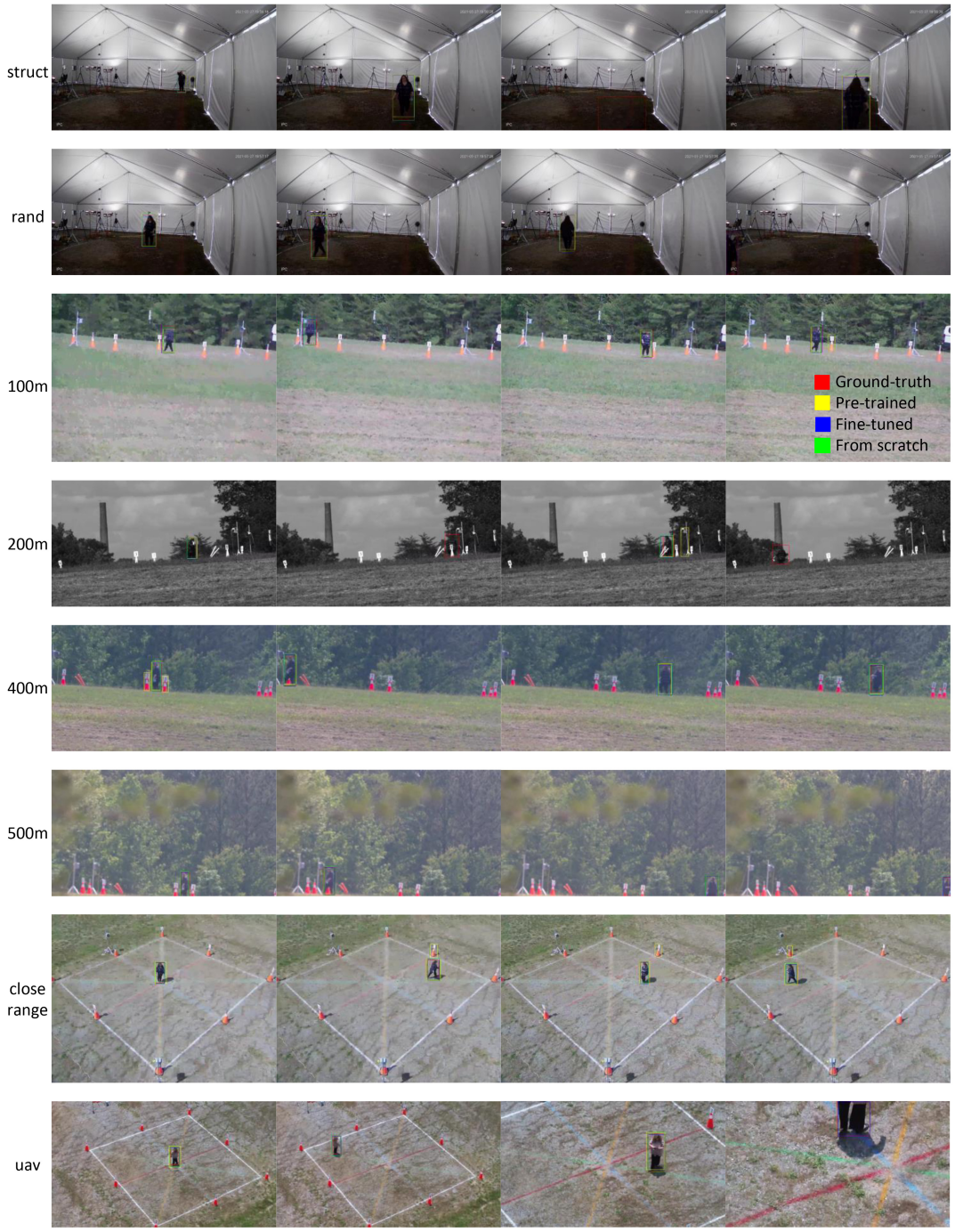


Fig. 5. Except for some extreme cases, models with BRIAR information can accurately detect the body on each dataset. This shows that our methods successfully address the challenges in processing the BRIAR dataset.

TABLE III
F1 SCORE COMPARISON IN OUTDOOR DATASETS OF BRIAR UNDER DIFFERENT IOU THRESHOLDS

	0.35				0.5				0.7			
model	**100m**	**200m**	**400m**	**500m**	**100m**	**200m**	**400m**	**500m**	**100m**	**200m**	**400m**	**500m**
PRE	93.19	88.53	87.01	82.09	93.02	88.41	86.77	81.97	90.55	85.06	82.57	76.57
FT	98.90	99.74	99.15	96.78	98.71	99.69	99.10	96.49	98.05	98.70	97.90	94.53
SCR	98.91	99.76	99.32	97.18	98.82	99.71	99.32	97.00	98.44	99.08	98.70	94.80
DET2	99.36	99.46	99.02	98.18	99.36	99.46	99.02	98.18	99.22	99.30	98.98	98.13

about learning body features. Finally, we notice that the fine-tuned model and the scratch model have significantly lower performance than Detectron2 at 500m, which further shows that at longer ranges, even if the model has the corresponding feature learning ability, if the body is too small or blurred, it is difficult to provide useful information to the model. However,

TABLE IV
F1 SCORE COMPARISON IN AERIAL DATASETS OF BRIAR UNDER DIFFERENT IOU THRESHOLDS

model	0.35		0.5		0.7	
	close range	UAV	close range	UAV	close range	UAV
PRE	94.75	70.45	94.69	69.48	91.81	61.14
FT	99.39	99.51	99.39	99.43	98.73	96.99
SCR	99.36	98.93	99.36	98.93	98.91	97.35
DET2	99.19	98.50	99.19	98.50	98.88	98.39

TABLE V
AVERAGE F1 SCORE COMPARISON IN ALL BRIAR DATASETS UNDER DIFFERENT IOU THRESHOLDS

model	0.35	0.5	0.7
PRE	89.57	89.46	86.10
FT	99.00	98.94	98.11
SCR	98.94	98.90	98.29
DET2	99.14	99.14	98.97

considering that the fine-tuned model achieves comparable performance with the scratched model and Detectron2 while using very little public dataset information, we claim that our method is efficient and accurate for various scenarios of BRIAR.

3) Detection in Aerial Data: The case of videos collected at altitude is relatively easier. Compared to outdoor results, all models have better performance on aerial video data (Table IV). This may be due to two reasons. First, in aerial datasets, subjects have less occlusions and incomplete bodies (Fig. 4, Fig. 5). Second, aerial cameras cover a large area of the ground. Subjects are less likely to fall out of camera range. These reasons make it possible for models to learn features more effectively in aerial situations. It can also be seen from the results at different IoU thresholds that all models are more stable on aerial than outdoor datasets. In summary, on aerial datasets, the fine-tuned model, scratched model and Detectron2 can all achieve $\sim$ 97+% of the F1 score.

We finally show the average results of all models in Table V. Note that the average results here are not a simple average of all previous results, but the F1 score of models tested on all BRIAR datasets. It can be seen that the fine-tuned model, scratched model and Detectron2 basically have similar performances. This suggests that our method does not require pre-training on massive public datasets, nor training on the BRIAR datasets for a long time from scratch, and can achieve an average of 99% of F1. Features learned by the fine-tuned model are sufficient to support downstream applications (such as verification, identification, recognition, etc.).

D. Visualizations of Detection Results

Fig. 5 shows the detection results of different models in consecutive video frames for each dataset. For many datasets, the pre-trained model has produced a large number of false negatives (such as the fan of images 1-2 in struct, the cone of images 2-4 in close range) and failures (such as in image 2 in 100m it did not detect the body). The fine-tuned model and the scratched model are more performant. However, for many corner cases, a degradation in performance is observed. For example, the subject image 3 in struct has hairs only on the back of the head and blends into the background rather well. At this time, except for the ground truth (red), the naked eye may not be able to accurately determine where the hair region is. Nevertheless, apart from these corner cases, it can be seen that in most indoor, outdoor and aerial scenarios, both the fine-tuned model and the scratch model can detect bodies well. This illustrates the effectiveness of our method.

Models that learn BRIAR information can produce extremely accurate detections in some cases. For example, for image 4 at 500m, the entire subject only occupies dozens of pixels in the corner of the image. For image 4 in UAV, the subject has only legs visible. However, the fine-tuned model can still accurately find the body in these cases, and the difference from ground truth is very small. In summary, using both the generality of the public dataset and the specificity of BRIAR, models can be trained quickly on both sides and learn a more robust representation. This may be very helpful to address the domain shift problem of representation.

E. Recognition and Identification Performance

Table VI presents a performance comparison of our model with DME [32] and CAL [43] on protocol 1. The primary metrics used for comparison are the accuracy of different ranks and the TAR under different FARs. Protocol 1 consists of a total of 185 subjects, including 100 distractors. The key differences among our models are as follows. Model 2 has a larger gallery size and a more difficult gallery than model 1, as it has more controlled images. Model 3, on the other hand, is trained on more challenging situations, including 270m, 370m, 490m, 600m, 800m, and 1000m. Additionally, model 3 is trained on face data simultaneously. Although face data may only include a partial body or even only a head, we included it in the training. It is evident that regardless of the model version, our models have achieved substantial performance improvements compared to CAL. Table VII presents a comparison of our results on the more challenging Protocol 2. This protocol consists of a total of 544 subjects, 444 of which are distractors. Despite the large number of distractors, our model achieves an accuracy of 73.30% at rank-20 and a TAR@1%FAR of 53.77%, demonstrating the effectiveness of our method for complex situations. Furthermore, model 2 outperforms model 1, while model 3 has achieved the highest performance, demonstrating the robustness of our model.

Tables VIII and IX summarize the performance comparison between state-of-the-art models and our models. Since FarSight [31] is a synthesized model of body, gait, and face, we only compare with their pure body version for a fair comparison. Under the same setting, our model has better performance than state-of-the-art methods. Considering that the backbone of both models is ResNet-50, this suggests that backbones decide the lower bound of the representation ability of bodies. On the other hand, BRIARNet has better rank-20 accuracy, while FarSight is slightly better in TAR. This suggests that different strategies focus on improving different aspects of representations, thereby moving to different upper bounds. A pure RGB feature representation (BRIARNet) tends to cluster similar identities, while 3D modeling (3DInvarReID [82] in

TABLE VI
PERFORMANCE COMPARISON OF OUR MODEL (**BRIARNET**) WITH STATE-OF-THE-ART ON PROTOCOL 1, BASED ON THE ACCURACY OF DIFFERENT RANKS AND THE TRUE ACCEPT RATE (TAR) AT DIFFERENT FALSE ACCEPT RATES (FAR). PROTOCOL 1 HAS 85 SUBJECTS AN 100 DISTRACTORS. SINCE BRIAR HAS BOTH FACE AND WHOLE BODY DATA, "WITH FACE" INDICATES WHETHER THE MODEL IS TRAINED WITH FACE DATA (Y) OR NOT (N). ALL RESULTS IN THIS PAPER CORRESPOND TO FACEINCLUDED PROTOCOL [16], [31]. THE "INPUT TYPE" INDICATES WHETHER THE MODEL USES IMAGES (I) AND/OR VIDEOS (V)

model	With Face	Input Type	Accuracy				TAR@FAR			
			Rank-1	Rank-5	Rank-10	Rank-20	0.01%	0.1%	1%	10%
DME [32]	N	V	24.23	51.49	64.02	74.37	0.02	0.02	2.15	22.97
CAL [43]	N	I	20.92	42.34	54.45	66.78	0.03	0.70	5.13	28.28
BRIARNet	N	I	36.66	66.86	78.98	87.59	2.79	11.13	35.69	71.28
	N	I + V	29.85	58.83	68.07	77.65	1.85	8.32	32.54	73.10
	Y	I + V	**46.67**	**73.75**	**82.51**	**90.36**	**4.13**	**15.97**	**49.26**	**86.78**

TABLE VII
PERFORMANCE COMPARISON OF OUR MODEL ON PROTOCOL 2. PROTOCOL 2 PRESENTS A MORE CHALLENGING TASK THAN PROTOCOL 1 DUE TO ITS 100 SUBJECTS AND 444 DISTRACTORS. NEVERTHELESS, OUR MODEL PERFORMED WELL, DEMONSTRATING ITS EFFECTIVENESS IN DIFFICULT SCENARIOS

model	With Face	Input Type	Accuracy				TAR@FAR			
			Rank-1	Rank-5	Rank-10	Rank-20	0.01%	0.1%	1%	10%
BRIARNet	N	I	12.18	24.27	25.72	37.17	2.24	7.94	22.55	43.64
	N	I + V	11.84	24.42	30.95	38.28	1.75	8.45	22.73	48.03
	Y	I + V	**34.84**	**55.93**	**64.70**	**73.30**	**6.61**	**24.90**	**53.77**	**83.03**

TABLE VIII
PERFORMANCE COMPARISON OF OUR MODELS WITH STATE-OF-THE-ART ON PROTOCOL 2. THE "BRS" COLUMN INDICATES WHICH DATASET WAS USED FOR TRAINING, I.E., DIFFERENT TRAINING SUBSETS DEFINED IN [16]

model	With Face	Input Type	BRS	Rank-20	1% FAR
CAL [43]	Y	I + V	1, 2	71.18	51.87
FarSight [31]	Y	I + V	1, 2	72.91	54.00
BRIARNet	N	I	1, 2	37.17	22.55
	N	I + V	1, 2	38.28	22.73
	Y	I + V	1, 2	73.30	53.77
	Y	I + V	1, 2, 3	**75.13**	**54.09**

TABLE IX
QUANTITATIVE COMPARISON BETWEEN PROPOSED METHOD, BRIARNET AND EXISTING APPROACHES ON BRIAR PROTOCOL 2. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

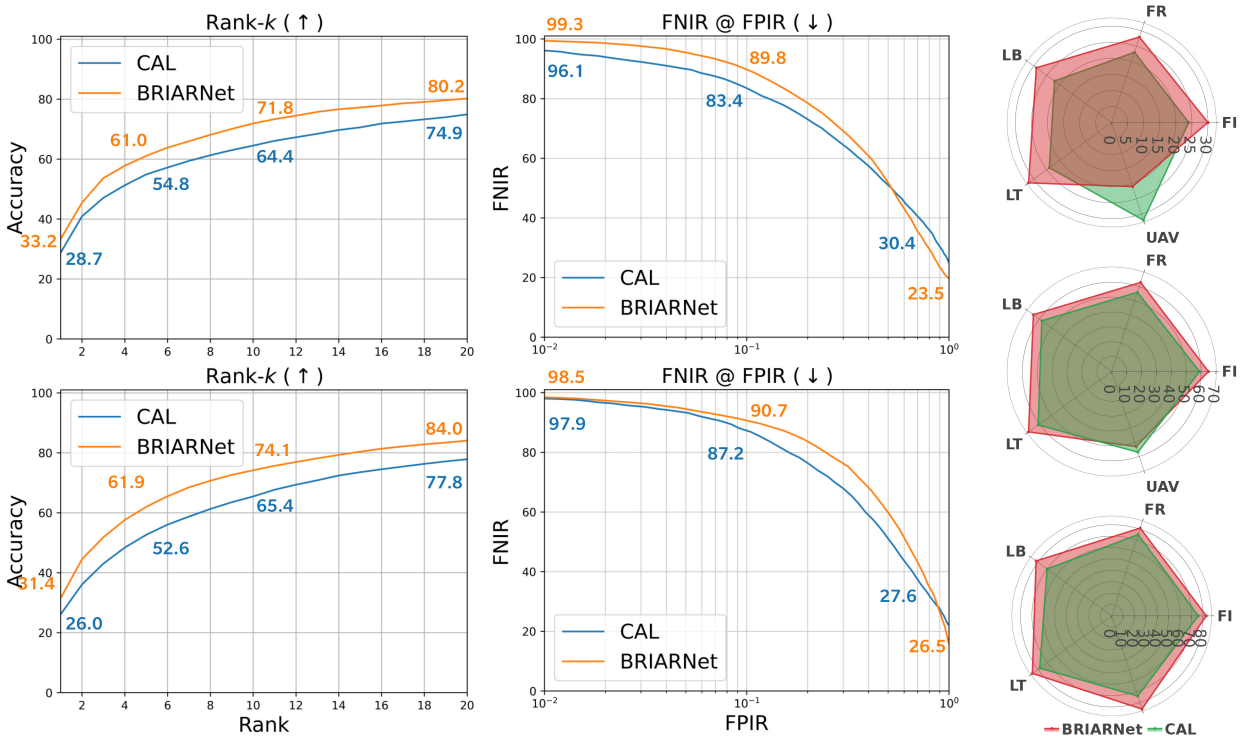
Methods (↓)	Rank				TAR @ x% FAR	
	1	5	10	20	$x = 1$	$x = 10$
CDNet [83]	7.05	19.82	30.27	41.66	15.80	54.55
CtF [84]	9.51	31.73	42.75	54.95	23.60	60.20
TransReID [85]	25.03	49.90	60.81	70.30	50.00	81.00
DC-Former [86]	27.98	50.46	62.28	72.94	49.38	84.51
PFD [87]	32.92	55.55	65.80	**75.73**	47.97	71.22
CAL [43]	19.72	43.64	54.02	64.94	38.50	74.68
BRIARNet	**34.84**	**56.95**	**66.33**	75.13	**54.09**	**89.03**

FarSight) tends to recognize the same identity. Besides, adding more data improves all metrics. Compared with our models from Tables VI to VIII, models trained with face images, videos, and various scenarios are better than models trained without them. Among them, face information is essential for good ReID.

To the best of our knowledge, there has been no comprehensive public report on whole-body recognition results for Protocol 3. In response, we present an evaluation of BRIARNet's performance on Protocol 3. Protocol 3 comprises two galleries: Gallery 1 consists of 134 subjects and 351 distractors, while Gallery 2 contains 130 subjects and 351 distractors. In Fig. 6, we display the performance curves of BRIARNet for both galleries. The FNIR denotes the fraction of failed mate searches that exceed a predefined threshold [88], serving as a metric for assessing open-set recognition performance. Specifically, on gallery 1, BRIARNet reaches 33.2%/61.0%/71.8%/80.2% rank-1/5/10/20 accuracy, which is higher than the SOTA method by 4.5%/6.2%/7.4%/5.3%. It is evident that BRIARNet's performance on the CMC curve remains largely unaffected, and in some cases, it even improves. Furthermore, even when confronted with a substantial number of distractors, BRIARNet demonstrates commendable open-set recognition capabilities (FNIR is $\sim$90% when FPIR $>$ 1%). In summary, our method has comparable performance to the SOTA method on most challenging open-set evaluation, and outperforms it significantly on rank-k accuracy.

We aim to assess the model's performance across various signature sets (sigsets), as illustrated in Fig. 6b. Five distinct sigsets available for extensive evaluations. Firstly, images are classified into Face Included (FI) and Face Restricted (FR) based on the visibility of the front face. Next, the BRIAR dataset exhibits natural grayscale variations and physical turbulence across a wide range of distances, presenting sigsets like Long-range Body (LB) and Long-range Turbulence (LT). Images captured from high altitudes by drones constitute the UAV sigset.

The evaluation focuses on rank-1, 10, and 20 for three sigsets. BRIARNet consistently outperforms CAL across most sigsets in rank-1. In FI, FR, LB, and LT, BRIARNet surpasses CAL by more than 5%, highlighting its robustness in scenes without faces, over long distances, and in the presence of severe turbulence. Notably, CAL leads BRIARNet in UAV, suggesting its suitability for recognizing high-altitude scenes. However, in rank-10 and rank-20 sigsets, BRIARNet maintains its lead in the first four sigsets while closing the gap in



(a) Protocol 3 - Gallery 1 and Gallery 2 results

(b) Sigset results

Fig. 6. (a) The performance curves of BRIARNet and CAL for both galleries on protocol 3. BRIARNet have >= 80% rank-20 accuracy, which is even higher than protocol 2's results. Besides, we also provide results of FNIR vs FPIR here as reference for other works in open-set recognition setting. (b) Sigset evaluation of two methods. FI, FR, LB, LT, UAV represent Face Included [16], [31], Face Restricted [16], [31], Long-range Body [16], Long-range Turbulence [16], Unmanned Aerial Vehicle [16], respectively. These results show the robustness of BRIARNet.

TABLE X
COMPARISON OF MODEL SIZE (IN MILLIONS) AND INFERENCE TIME (IN MILLISECONDS) ACROSS DIFFERENT METHODS FOR A IMAGE OF SIZE 256×128 ON A NVIDIA A5500 GPU

Methods (→)	TransReID [85]	CDNet [83]	CtF [84]	DC-Former [86]	PFD [87]	CAL [42]	Ours
Params	62.1	2.6	24.3	49.5	162	23.5	24.7
Time	23.2	21.2	40.9	20.6	28.4	29.9	38.2

UAV and eventually surpassing CAL by more than 10%. This underscores the overall effectiveness and robustness of BRIARNet across diverse scenarios. Also, in Table X, we provide the inference time and model size (in number of parameters) comparisons with other existing methods.

While the inference time (Table X) for our method is somewhat higher than others, such as DC-Former and CDNet, the notable improvement in recognition performance justifies this trade-off. Faster methods exhibit significantly lower accuracy, highlighting that BRIARNet offers a more favorable balance between speed and performance. The additional computational time in our method stems from steps that enhance robustness and accuracy, especially in challenging scenarios. Our method is specifically designed to handle complexities such as distances up to 500m, large pitch angles, severe turbulence, and varying environmental conditions – areas where faster models often fail. Furthermore, BRIARNet achieves superior recognition rates (as shown in Tables VIII and IX) and demonstrates robustness across diverse datasets, offering a compelling trade-off between speed and performance quality. We believe this trade-off is reasonable in practical applications where accuracy is paramount.

V. CONCLUSION

In this paper, we present the results of an end-to-end system for body detection and identification on real-world datasets, including the BRIAR dataset, under various altitudes and ranges. Our method generates features from public datasets and BRIAR, allowing our models to achieve high performance without exhaustive training on large datasets. By pre-training on a small number of public datasets and fine-tuning on BRIAR, our models can achieve 98% F1 score within 10 epochs. Additionally, our model achieved an accuracy of 75.13% and a TAR@1%FAR of 54.09%, outperforming state-of-the-art recognition and identification models. Experimental results demonstrate that models obtained through fine-tuning can maintain robust performance under different altitudes, ranges, environments, and actions.

ACKNOWLEDGMENT

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100005]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U. S. Government. The US. Government is authorized to reproduce

and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. The authors thank Drs. Josh Gleason, Matt Meyn, Nathan Shnidman, and Soraya Stevens for helpful discussions.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[5] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[7] W. Liu et al., "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[8] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[10] M. A. Khan, M. Mittal, L. M. Goyal, and S. Roy, "A deep survey on supervised learning based human detection and activity classification methods," *Multimedia Tools Appl.*, vol. 80, no. 18, pp. 27867–27923, 2021.

[11] S.-Z. Su, Z.-H. Liu, S.-P. Xu, S.-Z. Li, and R. Ji, "Sparse auto-encoder based feature learning for human body detection in depth image," *Signal Process.*, vol. 112, pp. 43–52, Jul. 2015.

[12] P. Doherty and P. Rudol, "A UAV search and rescue scenario with human body detection and geolocalization," in *Proc. Aust. Joint Conf. Artif. Intell.*, 2007, pp. 1–13.

[13] P. Rudol and P. Doherty, "Human body detection and geolocalization for UAV search and rescue missions using color and thermal imagery," in *Proc. IEEE Aerosp. Conf.*, 2008, pp. 1–8.

[14] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, 2017.

[15] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[16] D. Cornett, III et al., "Expanding accurate person recognition to new altitudes and ranges: The BRIAR dataset," 2022, *arXiv:2211.01917*.

[17] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Trans. Image Process.*, vol. 22, pp. 363–376, 2012.

[18] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 428–441.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, 2005, pp. 886–893.

[20] K. B. Bhangale and R. Shekokar, "Human body detection in static images using hog & piecewise linear SVM," *Int. J. Innov. Res. Develop.*, vol. 3, no. 6, pp. 179–184, 2014.

[21] H. Beiping and Z. Wen, "Fast human detection using motion detection and histogram of oriented gradients," *J. Comput.*, vol. 6, no. 8, pp. 1597–1604, 2011.

[22] R. Tong, D. Xie, and M. Tang, "Upper body human detection and segmentation in low contrast video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 9, pp. 1502–1509, Sep. 2013.

[23] D. Li, L. Xu, E. D. Goodman, Y. Xu, and Y. Wu, "Integrating a statistical background-foreground extraction algorithm and SVM classifier for pedestrian detection and tracking," *Integr. Comput.-Aided Eng.*, vol. 20, no. 3, pp. 201–216, 2013.

[24] Y. Kim and T. Moon, "Human detection and activity classification based on micro-doppler signatures using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 8–12, Jan. 2016.

[25] X. Chen, K. Henrickson, and Y. Wang, "Kinect-based pedestrian detection for crowded scenes," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 31, no. 3, pp. 229–240, 2016.

[26] W. Ouyang, X. Zeng, and X. Wang, "Partial occlusion handling in pedestrian detection with a deep model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 11, pp. 2123–2137, Nov. 2016.

[27] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 443–457.

[28] J. Li, Y. Wu, J. Zhao, L. Guan, C. Ye, and T. Yang, "Pedestrian detection with dilated convolution, region proposal network and boosted decision trees," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2017, pp. 4052–4057.

[29] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2017, pp. 953–961.

[30] S. Wu, H.-S. Wong, and S. Wang, "Variant semiboost for improving human detection in application scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 7, pp. 1595–1608, Jul. 2018.

[31] F. Liu et al., "FarSight: A physics-driven whole-body biometric system at large distance and altitude," 2023, *arXiv:2306.17206*.

[32] Y. Guo, C. Peng, C. P. Lau, and R. Chellappa, "Multi-modal human authentication using silhouettes, gait and RGB," in *Proc. IEEE 17th Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2023, pp. 1–7.

[33] G. Seguin, K. Alahari, J. Sivic, and I. Laptev, "Pose estimation and segmentation of multiple people in stereoscopic movies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1643–1655, Aug. 2015.

[34] S. Kamal and A. Jalal, "A hybrid feature extraction approach for human detection, tracking and activity recognition using depth sensors," *Arab. J. Sci. Eng.*, vol. 41, no. 3, pp. 1043–1051, 2016.

[35] M. H. Khan, K. Shirahama, M. S. Farid, and M. Grzegorzek, "Multiple human detection in depth images," in *Proc. IEEE 18th Int. Workshop Multimedia Signal Process. (MMSP)*, 2016, pp. 1–6.

[36] L. Qiang, W. Zhang, L. Hongliang, and K. N. Ngan, "Hybrid human detection and recognition in surveillance," *Neurocomputing*, vol. 194, pp. 10–23, Jun. 2016.

[37] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2018.

[38] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "FastReID: A PyTorch toolbox for general instance re-identification," 2020, *arXiv:2006.02631*.

[39] H. Luo et al., "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2597–2609, Oct. 2020.

[40] N. Martinel, G. L. Foresti, and C. Micheloni, "Aggregating deep pyramidal representations for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1–11.

[41] M. Wieczorek, B. Rychalska, and J. Dąbrowski, "On the unreasonable effectiveness of centroids in image retrieval," in *Proc. 28th Int. Conf. Neural Inf. Process.*, 2021, pp. 212–223.

[42] C. Eom, G. Lee, J. Lee, and B. Ham, "Video-based person re-identification with spatial and temporal memory networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12036–12045.

[43] X. Gu, H. Chang, B. Ma, S. Bai, S. Shan, and X. Chen, "Clothes-changing person re-identification with RGB modality only," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1060–1069.

[44] T. He, X. Jin, X. Shen, J. Huang, Z. Chen, and X.-S. Hua, "Dense interaction learning for video-based person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1490–1501.

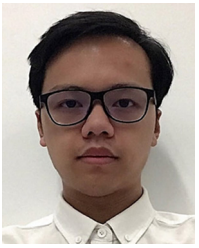
[45] A. Nambiar, A. Bernardino, and J. C. Nascimento, "Gait-based person re-identification: A survey," *ACM Comput. Surveys*, vol. 52, no. 2, pp. 1–34, 2019.

[46] X. Gu, H. Chang, B. Ma, H. Zhang, and X. Chen, "Appearance-preserving 3D convolution for video-based person re-identification," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 2020, pp. 228–243.

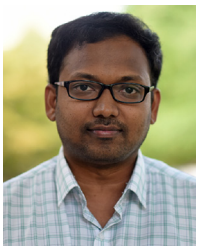
[47] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Temporal complementary learning for video person re-identification," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 2020, pp. 388–405.

[48] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "VRSTC: Occlusion-free video person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7183–7192.

[49] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "IAUnet: Global context-aware feature learning for person reidentification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 10, pp. 4460–4474, Oct. 2021.
[50] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Proc. 13th Eur. Conf. Comput. Vis.*, Zurich, Switzerland, 2014, pp. 688–703.
[51] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 11165–11172.
[52] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 608–617.
[53] Y. Ge et al., "FD-GAN: Pose-guided feature distilling GAN for robust person re-identification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–12.
[54] X. Qian et al., "Pose-normalized image generation for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 650–667.
[55] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3960–3969.
[56] X. Jin, C. Lan, W. Zeng, G. Wei, and Z. Chen, "Semantics-aligned representation learning for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 11173–11180.
[57] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 667–676.
[58] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7073–7082.
[59] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 542–551.
[60] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4678–4686.
[61] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2018, pp. 1–6.
[62] Z. Zhang et al., "Gait recognition via disentangled representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4710–4719.
[63] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2138–2147.
[64] Y. Huang, Q. Wu, J. Xu, and Y. Zhong, "Celebrities-ReID: A benchmark for clothes variation in long-term person re-identification," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2019, pp. 1–8.
[65] Y. Huang, J. Xu, Q. Wu, Y. Zhong, P. Zhang, and Z. Zhang, "Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3459–3471, Oct. 2020.
[66] X. Qian et al., "Long-term cloth-changing person re-identification," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 1–17.
[67] P. Hong, T. Wu, A. Wu, X. Han, and W.-S. Zheng, "Fine-grained shape-appearance mutual learning for cloth-changing person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10513–10522.
[68] X. Jin et al., "Cloth-changing person re-identification from a single image with gait prediction and regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14278–14287.
[69] L. Fan, T. Li, R. Fang, R. Hristov, Y. Yuan, and D. Katabi, "Learning longterm representations for person re-identification using radio signals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10699–10709.
[70] Q. Yang, A. Wu, and W.-S. Zheng, "Person re-identification by contour sketch under moderate clothing change," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2029–2046, Jun. 2021.
[71] J. Chen et al., "Learning 3D shape feature for texture-insensitive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8146–8155.
[72] Z. Gao, S. Wei, W. Guan, L. Zhu, M. Wang, and S. Chen, "Identity-guided collaborative learning for cloth-changing person reidentification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 2819–2837, May 2024.
[73] K. Han, S. Gong, Y. Huang, L. Wang, and T. Tan, "Clothing-change feature augmentation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22066–22075.
[74] Z. Yang, M. Lin, X. Zhong, Y. Wu, and Z. Wang, "Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1472–1481.
[75] S. Yang, B. Kang, and Y. Lee, "Sampling agnostic feature representation for long-term person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 6412–6423, 2022.
[76] V. D. Nguyen, K. Khaldi, D. Nguyen, P. Mantini, and S. Shah, "Contrastive viewpoint-aware shape learning for long-term person re-identification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 1041–1049.
[77] Y. Huang et al., "Meta clothing status calibration for long-term person re-identification," *IEEE Trans. Image Process.*, vol. 33, pp. 2334–2346, 2024.
[78] D. Arkushin, B. Cohen, S. Peleg, and O. Fried, "Reface: Improving clothes-changing re-identification with face features," 2022, *arXiv:2211.13807*.
[79] J. Byrne, G. Castanon, Z. Li, and G. Ettinger, "Fine-grained activities of people worldwide," 2022, *arXiv:2207.05182*.
[80] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.
[81] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. "Detectron2." 2019. [Online]. Available: https://github.com/facebookresearch/detectron2
[82] F. Liu, M. Kim, Z. Gu, A. Jain, and X. Liu, "Learning clothing and pose invariant 3D shape representation for long-term person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 19617–19626.
[83] H. Li, G. Wu, and W.-S. Zheng, "Combined depth space based architecture search for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6729–6738.
[84] G. Wang, S. Gong, J. Cheng, and Z. Hou, "Faster person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 275–292.
[85] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15013–15022.
[86] W. Li et al., "DC-Former: Diverse and compact transformer for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 2, 2023, pp. 1415–1423.
[87] T. Wang, H. Liu, P. Song, T. Guo, and W. Shi, "Pose-guided feature disentangling for occluded person re-identification based on transformer," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 2540–2549.
[88] P. Grother, M. Ngan, and K. Hanaoka, "Face recognition vendor test (FVRT): Part 3, demographic effects," U.S. Dept. Commerce, Nat. Inst. Stand. Technol., Gaithersburg, MD, USA, Rep. NISTIR 8280, 2019.



Siyuan Huang received the first M.S. degree in engineering from Johns Hopkins University and the second M.S. degree in computer science from George Washington University. He is currently pursuing the Ph.D. degree in electrical and computing engineering with Johns Hopkins University. He was an Associate Researcher with the Alternative Computing Group, National Institute of Standards and Technology, and a Graduate Research Assistant with the Tsinghua Laboratory of Brain and Intelligence, Tsinghua University. He works on person re-identification, body detection, recognition, and identification on large biometric person datasets.



Ram Prabhakar Kathirvel received the master's degree in electrical engineering from the National Institute of Technology Rourkela and the Ph.D. degree in computer vision from the Indian Institute of Science, Bengaluru, in 2022. He is an Assistant Research Scientist with Johns Hopkins University. His research interests include computer vision, computational photography, high dynamic range imaging, and remote sensing.

Yuxiang Guo (Graduate Student Member, IEEE) received the B.Eng. degree in electrical engineering from the University of Electronic Science and Technology of China and the University of Glasgow in 2019, and the M.Sc. degree in electrical engineering from Northwestern University in 2021. He is currently pursuing the Ph.D. degree in electrical and computing engineering with Johns Hopkins University. His research interests include computer vision, deep learning, and human-centered video-based identification.

Chun Pong Lau received the B.Sc. and M.Phil. degrees in mathematics from The Chinese University of Hong Kong in 2016 and 2018, respectively, the M.Sc. degree in applied mathematics from the University of Maryland in 2020, and the Ph.D. degree in computer science from Johns Hopkins University in 2021. He is currently an Assistant Professor with the School of Data Science, City University of Hong Kong. From 2022 to 2023, he was a Postdoctoral Fellow with the Mathematical Institute for Data Science, Johns Hopkins University. His research interests lie in developing reliable and robust computer vision algorithms, including atmospheric turbulence mitigation, adversarial robustness, biometrics at severe conditions, and generative AI.

Rama Chellappa (Life Fellow, IEEE) is a Bloomberg Distinguished Professor with the Department of Electrical and Computer Engineering (Whiting School of Engineering) and Biomedical Engineering (School of Medicine), Johns Hopkins University (JHU). At JHU, he is serving as an Interim Co-Director of the Data Science and Artificial Intelligence Institute and is also affiliated with CIS, CLSP, IAA, and MINDS. He holds a non-tenured position as a College Park Professor with the ECE Department, UMD. He holds nine patents. His research interests are in computer vision, pattern recognition, machine learning, and artificial intelligence. He received the 2012 K. S. Fu Prize from the International Association of Pattern Recognition. He is a recipient of the Society, Technical Achievement, and Meritorious Service Awards from the IEEE Signal Processing Society, the Technical Achievement and Meritorious Service Awards from the IEEE Computer Society, and the Inaugural Leadership Award from the IEEE Biometrics Council. He received the 2020 IEEE Jack S. Kilby Medal for Signal Processing, the 2023 IEEE Computer Society PAMI Distinguished Researcher Award, and the 2024 Edwin H. Land Medal from Optica (formerly, Optical Society of America). He is a member of the National Academy of Engineering and a Foreign Fellow of the Indian National Academy of Engineering. He is a Fellow of AAAI, AAAS, ACM, AIMBE, IAPR, NAI, OSA, and the Washington Academy of Sciences.